Journal of Biomedical Informatics 59 (2016) 258–275

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Computing semantic similarity between biomedical concepts using new information content approach

Mohamed Ben Aouicha, Mohamed Ali Hadj Taieb *

Multimedia InfoRmation system and Advanced Computing Laboratory, Sfax University, 3021, Tunisia

ARTICLE INFO

Article history:
Received 11 June 2015
Revised 6 November 2015
Accepted 12 December 2015
Available online 17 December 2015

Keywords:
Semantic similarity
Information content
DAG topological parameters
MeSH
Biomedicine

ABSTRACT

The exploitation of heterogeneous clinical sources and healthcare records is fundamental in clinical and translational research. The determination of semantic similarity between word pairs is an important component of text understanding that enables the processing and structuring of textual resources. Some of these measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data or from medical ontologies such as MeSH. This study focuses on Information Content (IC) based measures that exploit the topological parameters of the taxonomy to express the semantics of a concept. A new intrinsic IC computing method based on the taxonomical parameters of the ancestors' subgraph is then assigned to a biomedical concept into the "*is a*" hierarchy. Moreover, we present a study of the topological parameters through the MeSH taxonomy. This study treats the semantic interpretation and the different ways of expressing the parameters of depth and the descendants' subgraph. Using MeSH as an input ontology, the accuracy of our proposal is evaluated and compared against other IC-based measures according to several widely-used benchmarks of biomedical terms. The correlation between the results obtained for the evaluated measure using the proposed approach and those from the ratings of human' experts shows that our proposal outperforms the previous measures.

1. Introduction

Clinicians are confronted with increasing amounts of medical data from multiple sources housed in electronic format. The huge amounts of clinical and scientific documents in digital libraries and the digitized records assigned to patient health are valuable resources for clinical and translational research. Translational research includes medical information on patient health that comes from various sources and systems, including empirical observations, visits, and worksheet. The provided information is often heterogeneous and unprocessed. There is increasing interest in recent research in the search for variable strategies to manage and process this huge flow of data. The literature indicates that semantic technology offers promising opportunities for the development of efficient approaches to the interpretation of data from multiple origins and for determining the relationship between them.

The estimation of the semantic similarity between words is one of the major tools employed in semantic technology for text processing and understanding. It has been widely applied in several natural language processing tasks, such as word sense disambiguation [1,2], document categorization or clustering [3,4], word spelling correction [5], automatic language translation [4], ontology learning [6], and information retrieval [7,8].

In the biomedical field, the computation of the similarity between words can improve the performance of information retrieval from biomedical sources [8,9], integration of heterogeneous clinical data [10], automation of semantic grouping of clinical word pairs [11], and clustering of clinical models from local electronic health records [12].

Semantic similarity is a computational method used to identify and quantify likeness between words using the common characteristics shared between them. For example, *bronchitis* and *flu* are similar because they are both disorders of the respiratory system. The semantic similarity is based on the evaluation of the semantic evidence observed in a knowledge source (such as ontologies or domain corpora). According to the type of domain knowledge exploited, different families of functions can be identified: those based on the taxonomical structure of an ontology and those relying on the intrinsic Information Content (IC) of concepts [13–18].

These measures perform poorly with biomedical terms if they are exploited with general purpose knowledge [19], such as

* Corresponding author.
E-mail address: mohamedali.hadjtaieb@gmail.com (M.A. Hadj Taieb).

http://dx.doi.org/10.1016/j.jbi.2015.12.007
1532-0464/

WordNet[1] [20]. The problem with WordNet is not the quality of the semantic relations between the present biomedical concepts, but with its coverage capacity (only 25.1% of MeSH terms are covered in WordNet [19]). Therefore, there are a number of relevant biomedical ontologies, knowledge repositories and structured vocabularies that model and organize concepts in a comprehensive way. Well-known examples are MeSH (Medical Subject Headings) for indexing literature, the ICD taxonomy (International Classification of Diseases) for recording causes of death and diseases, and SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) the most comprehensive and precise clinical health terminology product, owned and distributed around the world by The International Health Terminology Standards Development Organisation (IHTSDO). Several similarity computation approaches have been compared using the biomedical knowledge source and evaluated over particular datasets or in the context of a concrete application, such as document clustering [3,21], assessment of similarity between words [8,22–24], and providing a useful basis for assessing the structure of terminological systems and the content of medical records [25].

In this paper, we first review and discuss the IC-based semantic similarity measures commonly referenced in the literature and provide details of their potential adaptation to the biomedical domain. We also analyze the taxonomic parameters used for the quantification of intrinsic information content of biomedical concepts to determine their semantic interpretations. In order to overcome some of the problems identified in this study, we present a new intrinsic IC computing method based on the exploitation of the ancestors' subgraph and the quantification of the specificity of each hypernym. Finally, the paper evaluates and compares the results obtained by our measure against those reported by other similarity functions when applied to the biomedical domain. The results show that our proposed method, coupled with Lin's similarity measure, displays a high level of correlation and outperforms other IC computing approaches.

The rest of the paper is organized as follows. Section 2 presents a survey about the IC-based semantic similarity measures, including the IC computing methods and the similarity measures. Section 3 provides a study of the topologic parameters extracted from the MeSH taxonomic knowledge resource for the computation of semantic similarity between biomedical concepts. Section 4 describes the new intrinsic IC-computing method of a biomedical concept based on its ancestors' subgraph and the taxonomic parameters. Section 5 reports on the evaluation and comparison of our approach against currently available ones using known benchmarks and the biomedical resource MeSH. The final section is devoted to presenting our conclusions and recommendations for future research.

2. Related works: information content-based semantic similarity measures

The measurement of semantic similarity based on Information Content (IC) was first introduced by Resnik [1]. The basic idea of IC is that general and abstract entities found in a discourse present less IC than more concrete and specialized ones. This principle is inspired from the work of Shannon [26]. The more probable a concept appears, the less information it conveys. In other words, specific words are more informative than general ones. IC-based semantic similarity measures [27–29] consist of two parts: the **computing IC method** and the **IC-based measure**. There are two ways for quantifying IC: the first exploits corpora, and the second, which is often described as *intrinsic*, uses topological parameters from the hierarchical knowledge structure: descendants (hyponyms), depth, leaves, and ancestors (hypernyms), for quantifying the IC of a concept. The terms "*hypernym/hyponym*", "*ancestors/descendants*" and "*subsumers*" are used as follows:

- *Hypernym/hyponym:* in the "*is a*" relation linking two concepts, such as "*Animal*" and "*Pet*", "*Animal*" is called hypernym of "*Pet*", and "*Pet*" is an hyponym of "*Animal*".
- *Ancestors/descendants:* ancestors of a concept pertaining to "*is a*" hierarchy refer to direct and indirect hypernyms. Descendants refer to direct and indirect hyponyms.
- Subsumer: a concept c_1 is a subsumer of c_2 if c_2 is a descendant of c_1.

IC-based similarity measures exploit the IC-values assigned to concepts c_1 and c_2 to provide the semantic similarity estimation between them. A complete survey of IC-based similarity measures is presented in the next paragraph.

2.1. Similarity measures exploiting the IC

Several semantic similarity measures, which are based on the exploitation of the information content, have been proposed. The similarity estimation between two concepts c_1 and c_2 is computed using their ICs and the IC of the Lowest Common Subsumer (LCS) which is extracted from the "*is a*" hierarchy. Some measures are presented in next paragraphs:

- *Resnik:* Guided by the idea that the similarity between a pair of concepts may be judged by "the amount of shared information", Resnik [1] defined the similarity between two concepts as the IC of their Lowest Common Subsumer *LCS(c_1,c_2)* as follows:

$$\mathrm{Sim}_{\mathrm{Res}}(c_1, c_2) = \mathrm{IC}(\mathrm{LCS}(c_1, c_2)) \tag{1}$$

- *Jiang-Conrath:* This approach subtracts the IC of the LCS from the sum of the IC of the individual concepts [30]. It provides the dissimilarity estimation between two terms, because the more different the terms are, the higher the difference between their ICs and the IC of their LCS will be. The dissimilarity measure is expressed as follows:

$$\mathrm{Dis}_{\mathrm{JC}}(c_1, c_2) = (\mathrm{IC}(c_1) + \mathrm{IC}(c_2)) - 2\mathrm{IC}(\mathrm{LCS}(c_1, c_2)) \tag{2}$$

- *Lin:* The similarity measure described by Lin [31] is defined as Dice coefficient:

$$\mathrm{Sim}_{\mathrm{Lin}}(c_1, c_2) = \frac{2 \times \mathrm{IC}(\mathrm{LCS}(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)} \tag{3}$$

- *Pirro:* He proposes a similarity measure [23] that is conceptually similar to the JC and Lin measures. However, it is based on the feature-based theory of similarity described by Tversky [32]. According to Tversky, the similarity between two concepts c_1 and c_2 is a function of the features common to c_1 and c_2, those in c_1 but not in c_2, and those in c_2 but not in c_1. The semantic similarity between concepts can be computed as an aggregation between the ICs of c_1, c_2, and their LCS:

$$\mathrm{Sim}_{\mathrm{tvr}}(c_1, c_2) = 3 \times \mathrm{IC}(\mathrm{LCS}(c_1, c_2)) - \mathrm{IC}(c_1) - \mathrm{IC}(c_2) \tag{4}$$

Finally, the measure is defined as follows:

$$\mathrm{Sim}_{\mathrm{P\&S}}(c_1, c_2) = \begin{cases} \mathrm{Sim}_{\mathrm{tvr}}(c_1, c_2) & \text{if } c_1 \neq c_2 \\ 1 & \text{if } c_1 = c_2 \end{cases} \tag{5}$$

- *Meng:* This measure [33] used Lin's measure. It increases monotonically with $\mathrm{Sim}_{\mathrm{Lin}}$ as follows:

$$\mathrm{Sim}_{\mathrm{Meng}}(c_1, c_2) = e^{\mathrm{Sim}_{\mathrm{Lin}}(c_1, c_2)} - 1 \tag{6}$$

[1] https://wordnet.princeton.edu/.

The measures detailed above are based on an IC that can be computed in two major methods, namely Corpora-based, where external resources are used, and intrinsic structure based, where knowledge structures are used. Both methods will be described in the following section.

2.2. IC computing methods

This section explains the methods of calculating the Information Content (IC) of a biomedical concept using external corpora or the topological parameters of the taxonomic knowledge resource. The IC values are exploited in next step to estimate the semantic similarity between two biomedical concepts.

2.2.1. Corpora-based IC quantification approach

The conventional way of measuring the IC of word meaning is to combine knowledge of their hierarchical structure from ontology with statistics on a large corpus. Resnik [1] estimated the frequencies of concepts in a taxonomy using noun frequencies from the Brown Corpus of American English [34]. The IC value of a concept c was then calculated by a negative log likelihood equation as follows:

$$\mathrm{IC}(c) = -\log(p(c)) \tag{7}$$

where p is the probability of subsuming c in a given corpus.

The corpus-based IC measure has several disadvantages. For example, the use of a general corpus for providing the ICs values of the concepts pertaining to a specific taxonomy like biomedical taxonomy leads to low performance. Therefore, the Brown corpus[2] has been used with the WordNet ontology as a corpus for general English words. As for the bioinformatics domain, clinical records and MEDLINE[3] (Medical Literature Analysis and Retrieval System Online) abstracts are used to compute the information content of biomedical concepts [35]. Another problem with corpus-based IC measures is that words are contained in a corpus whereas concepts are contained in ontology. Each polysemous word must be disambiguated to determine the correct sense which corresponds to a specific concept such in WordNet for example. This problem can be resolved by using annotated corpus which is rare and expensive. Also, an update performed on the corpus needs a recalculation of the probabilities assigned to the concepts pertaining to taxonomy.

For these reasons, several works have proposed the use of only the taxonomic structure without using an external corpus. These approaches are called *intrinsic information content*.

2.2.2. Intrinsic IC computation approach

Several works showed that the hierarchy structure of the knowledge resources can be exploited to extract the information content of a concept with no need for corpora analysis to compute the occurrence probability of a concept. This section presents the intrinsic IC computing methods following their chronological appearance in the literature.

Seco et al. [15] present a comprehensive IC intrinsic measurement that focuses only on the hierarchical structure of an ontology. The IC of a concept c depends on the concept which it subsumes. It is computed using the following equation:

$$\mathrm{IC}(c) = 1 - \frac{\log(|descendants(c)| + 1)}{\log(|T|)} \tag{8}$$

where *descendants(c)* refers to a function that returns the descendants of a given concept, and $|T|$ to a constant that represents the number of concepts in the knowledge resource T.

[2] http://khnt.hit.uib.no/icame/manuals/brown/index.htm.
[3] http://www.nlm.nih.gov/pubs/factsheets/medline.html.

Sebti et al. [14] use the hierarchical structure of the resource and implicitly includes the depth of a target concept. This method is based on the number of direct hyponyms of each concept pertaining to the initial path of the root till reaching the target concept.

In Fig. 1, the numbers on the left represent the number of direct hyponyms for each concept. For a better understanding of this method, Eq. (9) computes the IC of the concept *D018123*:

$$\mathrm{IC}(D018123) = -\mathrm{Log}\left(\frac{1}{16}\times\frac{1}{97}\times\frac{1}{39}\times\frac{1}{32}\times\frac{1}{16}\times\frac{1}{9}\times\frac{1}{18}\right) = 9.7007 \tag{9}$$

Zhou et al. [16] present a new approach to overcome the limitations associated with the *Seco* et al. method. Their method considers only the hyponyms of a given concept. In brief, with *Seco* method, the concepts having the same number of hyponyms have equally IC value (Eq. (8)) despite their different degrees of generality expressed by the depth. So, *Zhou* et al. [16] method was proposed to enhance descendants-based IC computation with the relative depth of the concept in the taxonomy, which was integrated in a formula with a tuning factor:

$$\mathrm{IC}(c) = k\left(1 - \frac{\log(|descendants(c)|+1)}{\log(|T|)}\right) + (1-k)\left(\frac{\log(depth(c))}{\log(max\_depth)}\right) \tag{10}$$

In addition to *descendants(c)* and $|T|$, which have the same meaning as in Eq. (8), *depth(c)* refers to the depth of the concept c in the taxonomy, and *max_depth* to the maximum depth of the taxonomy. The parameter k is a tuning factor that adjusts the weight of the two features used in the IC formula.

Sánchez et al. [17] followed another strategy and did not include the depth notion. They used the descendants through the leaves of the descendants' subgraph of a concept and integrated a novel parameter, *subsumers(c)*. They consider that the leaves are sufficient enough to describe and differentiate one concept from any other. Formally, they define the *leaves* and *subsumers* of a concept c as follows:

Leaves(c) = $\{l \in T / l \in hyponyms(c) \bigwedge l \text{ is a leaf}\}$, where T is the set of concepts of the taxonomy.
Subsumers(c) = $\{a \in T / c \leqslant a\}$, where $c \leqslant a$ means that c is a hierarchical specialization of a.

Following a similar principle in related works, they consider that concepts with several leaves in their descendants' subgraph are general (i.e., they have low IC) because they subsume the meaning of many important terms.

They proposed the following IC formula:

$$\mathrm{IC}(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|}+1}{max\_leaves+1}\right) \tag{11}$$

Meng et al. [13] present a formula that merges the principles used by Seco and Zhou. They have also changed the term $|descendants(c)|$ by another term to better express the hyponyms contribution in the IC of a concept. The novel term integrates the depth notion in the Seco formula.

Their method expressed IC as follows:

$$\mathrm{IC}(c) = \frac{\log(depth(c))}{\log(max\_depth)} \times \left(1 - \frac{\log(\sum_{a\in descendants(c)}\frac{1}{depth(a)}+1)}{\log(|T|)}\right) \tag{12}$$

For a given concept c, *depth(c)* refers to the depth of concept c in the taxonomy, *max_depth* to the maximum depth in the taxonomy, and $|T|$ to the number of concepts that exists in the taxonomy T.

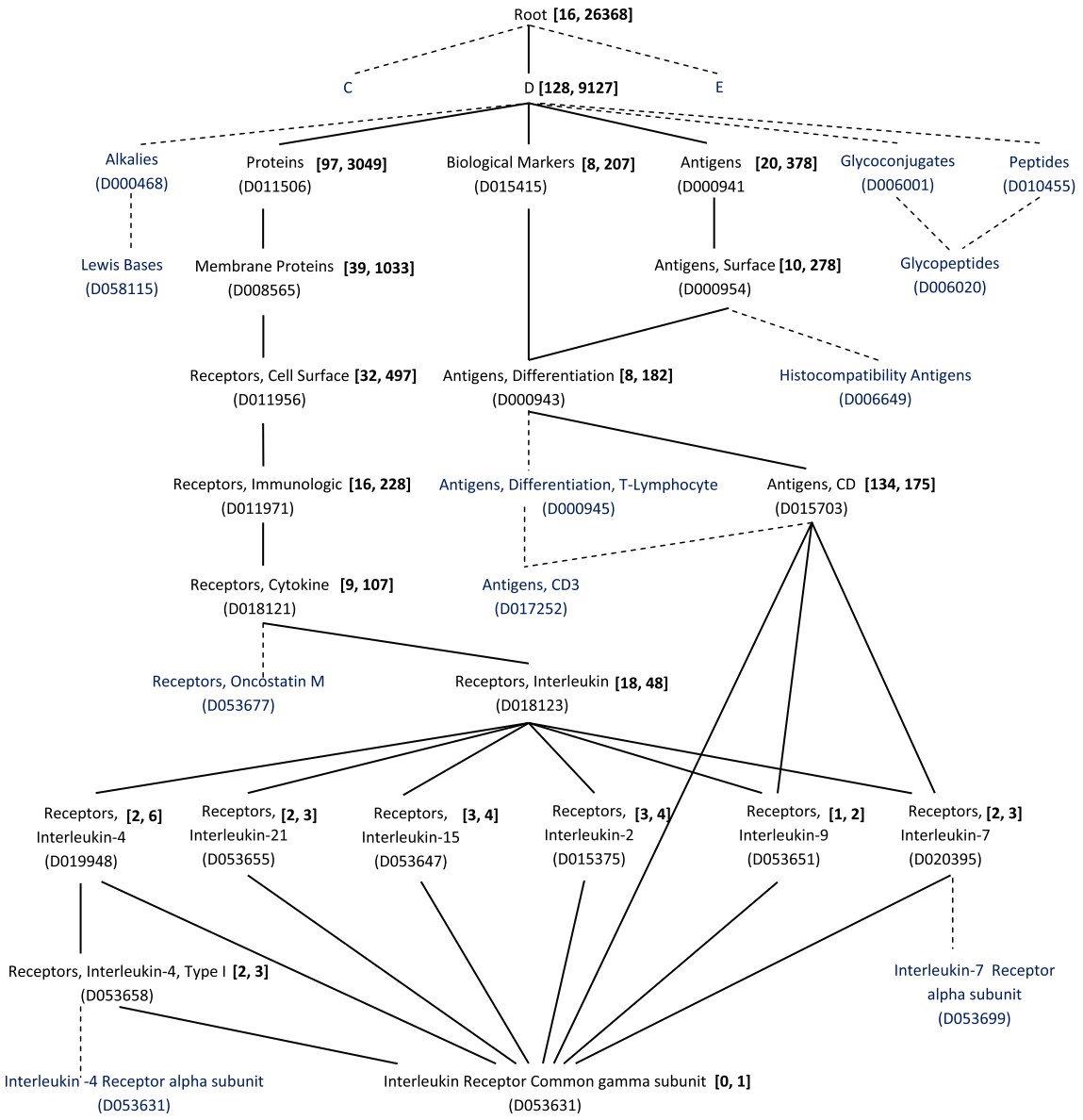

Fig. 1. An excerpt from the MeSH taxonomy. The information between brackets refers the [number of direct hyponyms, number of the descendants subsumed by the target concept including the concept].

Several studies have shown that the intrinsic computing methods provide better results than the ones based on the statistic analysis of corpora [36] when they are compared to small datasets such as RG65 [37] and MC30 [38] with WordNet. Despite that, the main limitation of the intrinsic IC approach is its limited coverage capacity. In fact, a concept that does not pertain to the "*is a*" hierarchy does not have an IC value. Furthermore, for some taxonomies, such as the hierarchy "*is a*" of WordNet, a word can be represented by a great number of senses (the word "head", for instance, has 33 senses), which can affect negatively the semantic similarity estimation due to the fact that the computing process will exploit rarely used senses. Moreover, the knowledge structure contains some concepts that do not exist in texts [39].

2.2.3. Summary

Table 1 shows that the IC-based similarity measures provide positive and/or negative values to refer the similarity or dissimilarity between biomedical concepts. The measures expressing the dissimilarity lead to negative correlations with the benchmarks annotated for semantic similarity purpose. Table 2 shows that most of the intrinsic IC computing methods exploit the depth and quantify the descendants' subgraph assigned to a target concept.

In the following section, we present a study related to the topological parameters of the MeSH "*is a*" hierarchy. These parameters are exploited in the intrinsic IC computing methods.

3. Topological parameters' study of MeSH taxonomy

Semantic similarity measures the likeness between concepts using information from predefined knowledge sources (such as ontologies or domain corpora) and a number of topological parameters related to the taxonomy (view the excerpt from MeSH in Fig. 1). Information content-based approach quantifies the similarity between concepts as a function of the Information Content (IC) that both concepts have in common in a given ontology.

In this section, we study the taxonomical parameters of the MeSH "*is a*" hierarchy that are exploited to design taxonomical

Table 1
IC-based similarity measures.

	Ref.	Positive	Negative	Similarity	Dissimilarity
IC-based semantic similarity measures					
Resnik	[1]	X		X	
Jiang and Conrath	[30]	X			X
Lin	[31]	X		X	
Pirro	[23]	X	X	X	
Meng and Gu	[33]	X		X	

Table 2
Intrinsic IC computing methods and taxonomical parameters.

	Ref.	Depth	Hyponyms	Leaves	Hypernyms
Intrinsic IC computing methods					
Seco et al.	[15]		X		
Sebti and Barfroush	[14]	X	X		
Zhou et al.	[16]	X	X		
Sanchez et al.	[17]			X	X
Meng et al.	[13]	X	X		

measures, such as the IC-based measures. The main parameters used in the literature are: descendants, depth, leaves, and ancestors. This study focuses on the semantic interpretation, dependencies, and probability distribution of those parameters in the taxonomy.

3.1. The depth

The depth of a concept is a parameter used in several measures for semantic similarity computing. The concept existing at the top of the taxonomy represents a general concept. The concept occurring at the bottom of the taxonomy represents a specific concept and is semantically richer.

The depth is an important element in determining the specificity of a concept because going down in the taxonomy from a parent to its descendant's leads to the propagation of features which will be enriched by some specificity.

The depth is commonly used to express the specificity of a concept inside a taxonomy, such as the MeSH taxonomy, which is considered as directed acyclic graph. It is used mainly as the longest path between a concept and its root (*depthmax*). In some cases, however, the depth is expressed as the shortest path between the target node and the root (*depthmin*). In their recent work [40], Wang and Hirst proposed a new method for representing the depth ($depth_{WH}$) using the distribution of the classic depth's definition as illustrated by Equation (13).

$$depth_{WH}(c) = \frac{\sum_{c' \in T} |\{c' : depth(c') \leqslant depth(c)\}|}{|T|} \tag{13}$$

In Eq. (13), $depth(c)$ represents the longest path between the node c (c representing a biomedical concept) and the root in a taxonomy T, and $|T|$ is the total number of nodes pertaining to the taxonomy T.

The transition from concept c_1 towards concept c_2 using the "*is a*" relation does not mean the passage from depth i to depth $i+1$. Thus, two concepts that are directly connected do not necessarily have successive depths in the taxonomy (for example, in Fig. 1, the concepts *"Interleukin Receptor Common gamma subunit" (D053631)* and *"Antigens, CD" (D015703)* are directly connected but the depthmax is equal to 10 for the first concept and to 5 for the second concept).

Fig. 2 shows the number of nodes in each depth in relation to the MeSH taxonomy. For the *depthmax*, it can be interpreted as the number of nodes per layer. Fig. 2 shows that the depth of the MeSH taxonomy[4] is 16. Fig. 2a and b shows that most of the nodes pertaining to the MeSH "*is a*" hierarchy have different *depthmax* and *depthmin* (38.84%). The maximum difference is found for the concept "*Philadelphia Chromosome*", with *depthmax* = 16 and *depthmin* = 5. This difference can be explained by the fact that the multiple inheritances are frequent in biomedical taxonomies such the one of MeSH. Also, the distribution of the $depth_{WH}$ is the same as for the *depthmax* (Fig. 2a) except that the depth of each node is represented by the probability of the *depthmax* in the taxonomy (Eq. (13)).

3.2. The descendants

The descendants subsumed by a concept c, noted as *descendants (c)*, includes the concept c. This parameter is also used, as a depth, to determine the generality/specificity feature. This idea corresponds well with the notion of Information Content (IC) because a general concept containing a large set of descendants (direct and indirect hyponyms) is considered as more probable. Then, the concept would become more probably represented by low information content according to the theory of information content presented by Shannon [26]. It does not, however, involve the case of leaf concepts, which have high information content values.

This parameter has, therefore, been used by Seco et al. [15] as another alternative strategy for computing IC independently from the corpus processing task to determine the probability of each concept pertaining to the taxonomy. Fig. 3 shows that the majority of biomedical concepts in the MeSH taxonomy are leaf nodes. So, they have a higher degree of specificity.

Among descendants, we can exploit only the leaf concepts based on the fact that the concepts having several leaves in their descendants' subgraph are qualified as general concepts since they subsume the meaning of many salient concepts. Leaves present equally specialized concepts that are completely disambiguated when compared to general concepts in the taxonomy, such as "*Proteins*" (*D011506*), and the leaf concept "*Interleukin Receptor Common gamma subunit*" (*D053631*). The descendants assigned to a biomedical concept can be replaced by only the leaves pertaining to this set. A concept, subsuming a considerable number of leaves, is considered as general concept. The leaves represent 66.75% of the whole nodes in the whole graph.

The descendants of a concept c form a subgraph that can be quantified in several ways to evolve into a semantic similarity model. As will be described below, the literature presents several methods for quantifying the descendants' subgraph formed by direct and indirect hyponyms (**QuantifiedDescendants**) of a given concept in the taxonomy.

Seco et al. [15] took the cardinality of the descendants set *Descendants(c)* (Eq. (14)). Some other works exploited only the leaf nodes *DescendantsLeaf(c)* [17] (Eq. (15)).

$$QuantifiedDescendants_1(c) = |Descendants(c)| \tag{14}$$

$$QuantifiedDescendants_2(c) = |DescendantsLeaf(c)| \tag{15}$$

Meng et al. [13] employed the depth to take the specificity of each concept pertaining to the set of descendants into consideration:

$$QuantifiedDescendants_3(c) = \sum_{c' \in Descendants(c)} \frac{1}{Depth(c')} \tag{16}$$

[4] The depth of the taxonomy is the maximum *depthmax* between the root and a node.

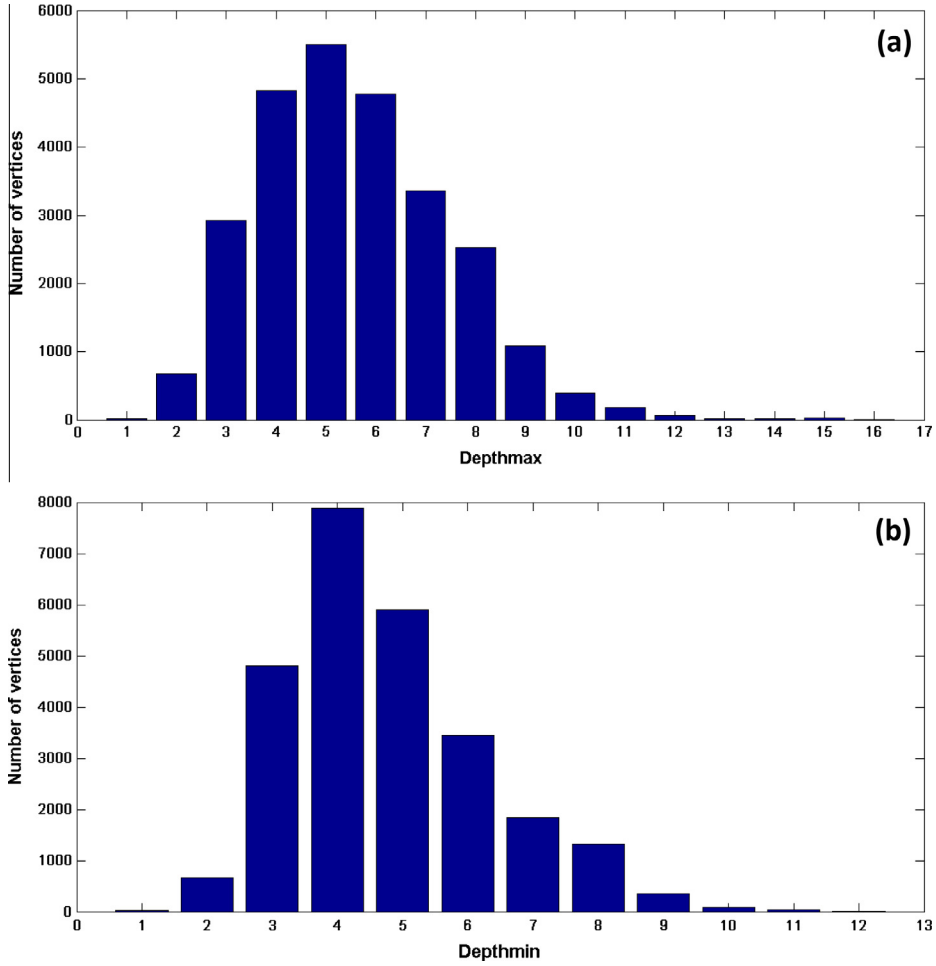

Fig. 2. The distribution of different depth-expressing ways through the taxonomy of MeSH.

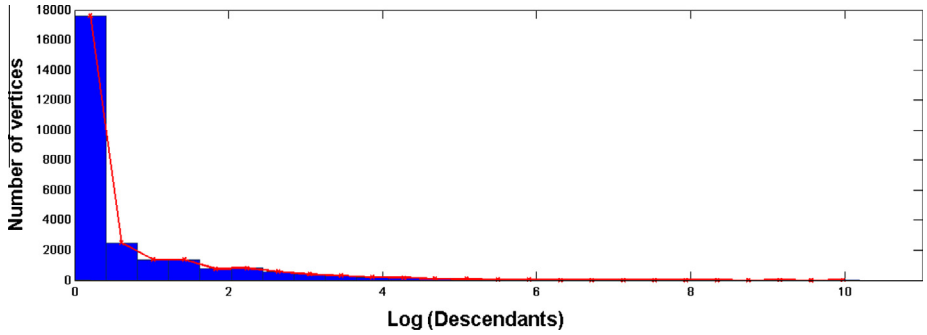

Fig. 3. The distribution of the descendants' number in a logarithmic scale through the MeSH taxonomy.

$$QuantifiedDescendants_4(c) = \sum_{c' \in Descendants(c)} \frac{1}{depth_{WH}(c')} \quad (17)$$

The depth is expressed in two ways: the *depthmax* (Eq. (16)) and the novel definition $depth_{WH}$ (Eq. (17)).

Hadj Taieb et al. [36] proposed the *QuantifiedDescendants(c)* method, which is based on the depth probability distribution over the taxonomy. This method proceeds as follows:

$$QuantifiedDescendants_5(c) = \sum_{c' \in \text{Descendants}(c)} \text{P}(\text{depth}(c')) \quad (18)$$

where *Descendants(c)* is the descendants set of the concept *c* and *depth(c)* represents the *depthmax* between a given concept *c* and the root. The depth probability *P(depth(c))* is then computed as follows:

$$P(depth(c)) = \frac{|\{c' \in T | depth(c') = depth(c)\}|}{N} \quad (19)$$

where *T* is the set of concepts pertaining to the MeSH taxonomy, and *N* is the cardinality of the set *T*.

3.3. Taxonomical ancestors meaning

Fig. 4 shows the number of biomedical concepts pertaining to the taxonomy having a specified number of ancestors (direct and indirect hypernyms). The richness of ancestors assigned to a

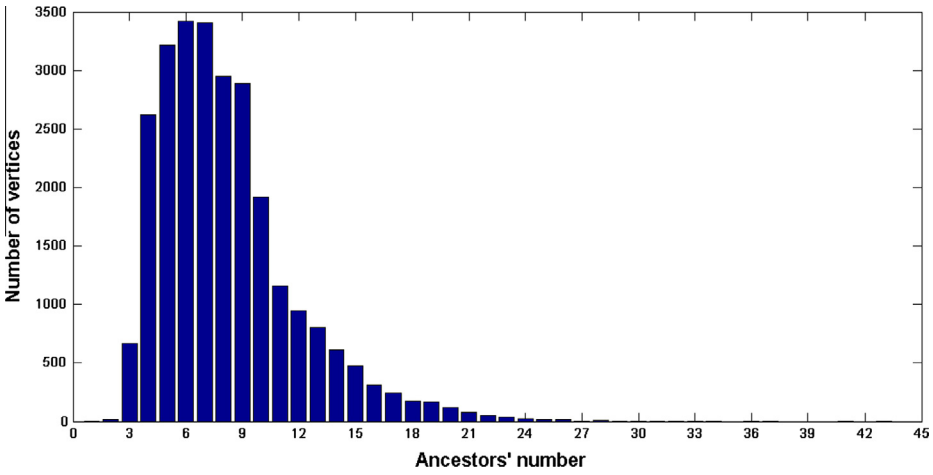

Fig. 4. The distribution of hypernyms' number through the MeSH taxonomy.

biomedical concept, as illustrated in Fig. 4, is due to the multiple inheritances. The ontologies modeling multiple inheritances may incorporate several direct subsumers per concept. In MeSH, 28.16% of the nodes in the taxonomy are with multiple inheritances according to the performed statistics on WordNet 3.0 (2.28%). The effect of multiple inheritances has been invoked in various fields at a deeper-level. In cases of multiple inheritances, some related works on ontology consider only the subsumer that defines the maximum of shared characteristics between the target concept pair.

However, when a concept inherits from several subsumers, it becomes more specific than another one inheriting from a unique subsumer. The number of subsumers cannot be considered as the depthmax parameter because 51.63% of the nodes in MeSH have an ancestors' number that is different from *depthmax*. Therefore, we focus on the ancestors' subgraph of a biomedical concept to express its semantics using the information content due to the frequency of the multiple inheritances in the MeSH knowledge source.

4. Proposed method: Ancestors' subGraph-based IC (AsGIC)

Our approach exploits the ontological structure seeking to achieve a better semantic understanding of a concept. As explained in the previous section, intrinsic IC computing methods use several topological parameters, including the depth, hyponyms, ancestors, and Lowest Common Subsumer. The ancestors' subgraph is exploited by some authors [41,42] to express the semantic similarity based only on the cardinality of ancestors' set. But, our proposal is to first to express the information content by weighting each concept pertaining to the ancestors' subgraph modeling the semantics of a biomedical concept. In MeSH, 28.16% of the nodes in MeSH have multiple inheritances. Therefore, the IC assigned to a concept is quantified using the contribution of each ancestor belonging to the subgraph. The two cornerstones of our intrinsic IC method are explained in the following paragraphs.

4.1. The ancestors' subgraph and the features' propagation

In the "*is a*" knowledge structure, a concept inherits the basic features from the ancestor concept and adds its own specific features to form its semantics. Accordingly, the meaning of a concept is an accumulation of the features coming from a higher ancestor to another less deep. Thus, a concept depends strongly on its hypernyms (direct parents) and ancestors. The IC of a biomedical concept is, therefore, modeled by the subgraph formed by the concept ancestors.

For example, in Fig. 1, the ancestors' subgraph assigned to the concept *D053631* is represented by the solid lines linking the biomedical concepts. It is formed by the concepts {*D053631*, *D020395*, *D053651*, *D015375*, *D053647*, *D053655*, *D019948*, *D018123*, *D018121*, *D015703*, *D011956*, *D000943*, *D000954*, *D000941*, *D008565*, *D011506*, *D015415*, *D*, *Root*}. So, the IC of the node *D053631* is quantified by expressing the contribution of each ancestor. This contribution is estimated by expressing the specificity of the biomedical concept in the hierarchical "*is a*" structure.

Fig. 5 shows the distribution of hypernyms' number through the MeSH taxonomy. It also demonstrates the importance of multiple inheritances in this taxonomy. So, the meaning of a concept is formed by the features of all its ancestors.

4.2. Quantifying the specificity using topological parameters

The contribution of each ancestor pertaining to the ancestors' subgraph is computed based on the specificity estimation. This specificity is calculated using the hypernyms of each concept. According to Fig. 1, the contribution of the ancestor *D000943* to the IC of *D053631* is calculated using its hypernyms *D000954* and *D015415*, and the contribution of *D053631* is computed using its hypernyms {*D053658*, *D019948*, *D053655*, *D053647*, *D0153775*, *D053651*, *D020395*, *D015703*}. The specificity of each hypernym is computed using the topological parameters: the depth and the descendants. The hypernyms assigned to a target concept have different depths. They, therefore, have different degrees of specificity. For example, the hypernyms of the concept *D053631* have the depths 7 (*D015703*), 8 (*D019948*, *D053655*, *D053647*, *D015375*, *D053651* and *D020395*) and 9 (*D053658*). The integration of the descendants' parameter in the specificity quantification process is expressed as the quantification of the overlapping part between the descendants' subgraph assigned to the hypernym and the one assigned to its concerned hyponym. This overlap can be computed in several ways because the descendants' subgraph can be expressed in different forms as explained in Section 3.

4.3. Ancestors' subGraph-based IC computing method

The proposed computing method *AsGIC* is a new intrinsic IC computing method of a biomedical concept in the MeSH taxonomy as

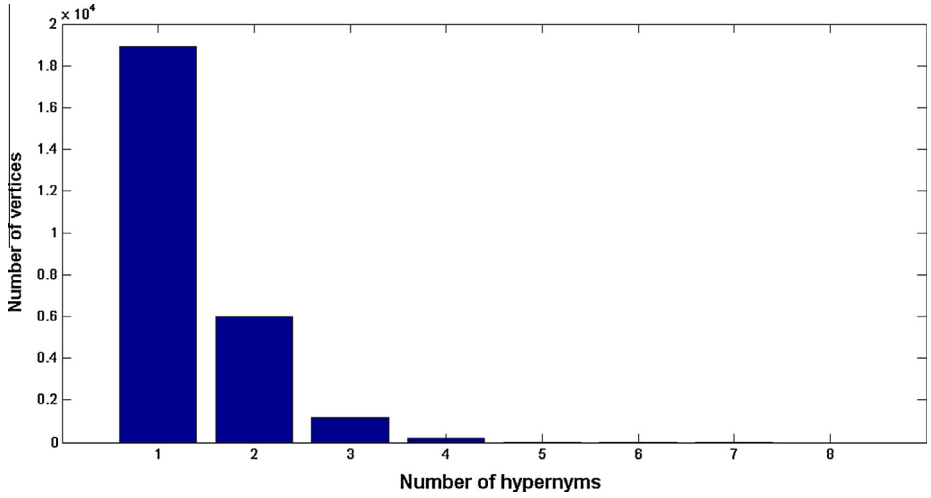

Fig. 5. The distribution of the direct hypernyms number through the MeSH taxonomy.

modeled by its assigned ancestors' subgraph (Fig. 1). The information content of a biomedical concept *BioCon* is computed as follows:

$$IC(BioCon) = \sum_{c \in Ancestors(BioCon)} AncestorSpecificity(c) \quad (20)$$

IC(BioCon) refers to the accumulation of the ancestors contributions pertaining to *Ancestors(BioCon)* and forming the ancestors' subgraph. The contribution of each ancestor is calculated according to its specificity quantified by the "*is a*" hierarchy.

Formally, we define the sets *Ancestors* and *Descendants* as follows:

Definition 1. The taxonomy "*is a*" is modeled as a Directed Acyclic Graph (DAG) G defined by $<V,E>$ where V refers the vertices' set and E refers the edges' set. For the vertices a and b pertaining to G, $Path(a,b)$ is a path connecting a to b. Also, we define $Depth(a)$ as the *depthmax* of the vertex a.

Definition 2. Let the concept subsumption ($\leqslant$) be a binary relation $\leqslant : T \times T$, with T being the set of concepts in the taxonomy, where $d \leqslant c$ means that d is a hierarchical specialization of c defined as:

$$d \leqslant c = \{\exists\, Path(c,d)/depth(c) \geqslant depth(d) \text{ with } (c,d) \in V^2\}$$

We define the set *Ancestors* of a concept d as follows:

$$Ancestors(d) = \{c \in C/d \leqslant c\}$$

And the set *Descendants* of a concept d as follows:

$$Descendants(d) = \{c \in C/c \leqslant d\}$$

For example, in relation with Fig. 1, $Ancestors(D000943)$ = {*D000943*, *D000954*, *D015415*, *D000941*, *D*, *Root*}.

Definition 3. Let the direct concept subsumption ($@\rightarrow$) be a binary relation $@\rightarrow : T \times T$, with T being the set on concepts in the taxonomy, where $d\ @\rightarrow\ c$ means that d is a hyponym of c defined as:

$$d@\rightarrow c = \{\exists\, Path(c,d)/|Path(c,d)| = 1 \text{ and } depth(c) \geqslant depth(d) \text{ with } (c,d) \in V^2\}$$

We define the set *Hyper*nyms of a concept d as:

$$Hypernyms(d) = \{c \in C/d@\rightarrow c\}$$

For example, in relation with Fig. 1, $Hypernyms(D000943)$= {*D000954*, *D015415*}.

The specificity of each ancestor $AncestorSpecificity(c)$ is calculated using the topological parameters: depth and descendants of its hypernyms as follows:

$$AncestorSpecificity(c) = \sum_{c' \in Hypernyms(c)} (Depth(c') \times HypoOverlap(c,c')) \quad (21)$$

where $Depth(c')$ represents the depth of the biomedical concept c'; it refers to the depth of the concept c in the taxonomy T and is expressed as the *depthmax*, *depthmin* or $depth_{WH}$ as explained in Section 2. $HypoOverlap(c,c')$ quantifies the overlapping part between the descendants' subgraphs of the concepts c and c'. $HypoOverlap(c,c')$ is expressed as follows:

$$HypoOverlap(c,c') = \frac{QuantifiedDescendants(c)}{QuantifiedDescendants(c')} \quad (22)$$

The function $QuantifiedDescendants(c)$ serves for quantifying the descendants' subgraph, which can be expressed using different ways from Eqs. (14)–(18). All these methods will be examined in the experiments section.

4.4. Study of AsGIC' topological parameters

In Figs. 3–5 presented above, we studied the distribution of topological parameters: the depth, hyponyms, and ancestors according to the MeSH taxonomy. In Fig. 6, we detailed the relation between the IC values computed with the proposed method and the topological parameters. For each biomedical concept *BioCon* pertaining to the MeSH taxonomy, the *IC(BioCon)* is computed using the *depthmax* and hyponyms cardinality (Eq. (14)). Fig. 6 shows the distribution of the $IC(BioCon)$ through the hypernyms number composing the ancestors' subgraph (Fig. 6(a)), the depth of the concerned concept (Fig. 6(b)), and the number of descendants subsumed by the concept *BioCon*. Fig. 6(a) demonstrates that for the same $|Ancestors(BioCon)|$ value, the proposed IC computing method provides different IC values according to the specificity of each ancestor. Moreover, when the numbers of concepts forming the ancestors' subgraph modeling the IC of the concept *BioCon* increases, the method does not automatically give a high IC value. The concepts marked in Fig. 6(a) have a considerable number of ancestors between 30 and 35, but their IC-values are weak.

Fig. 6(b) shows that deeper concepts have high IC-values due to the enrichment process by the features inherited and propagated from an ancestor to another. It can also be noted that, for the same

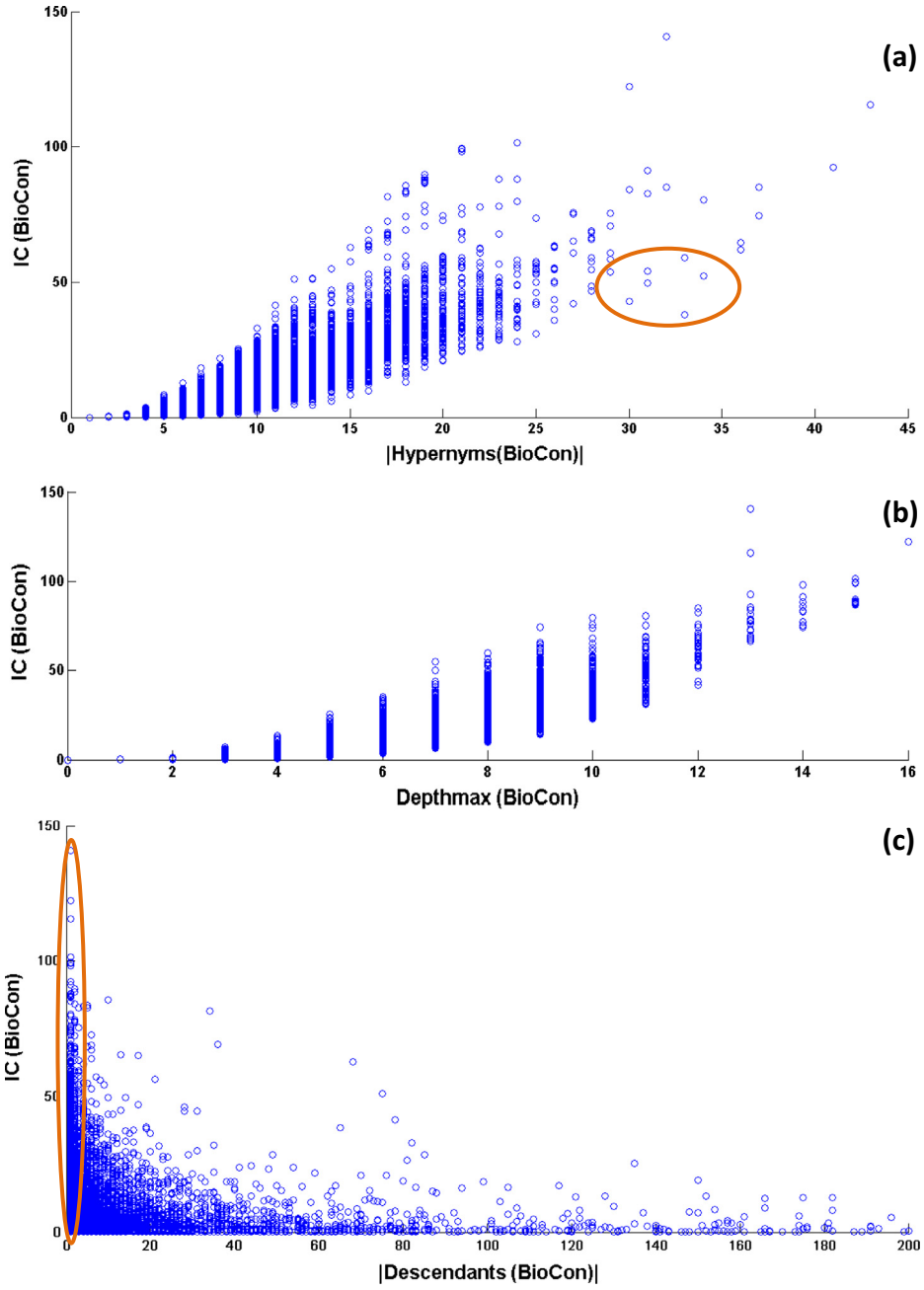

Fig. 6. Study of the IC-values distribution in relation with the ancestors (a), *depthmax* (b), and descendants (c).

depth, there exists a high scale for the IC values (for example, with *depthmax* = 9, IC-values $\in$ [14.2303, 74.4168]. Fig. 6(c) demonstrates that our proposal provides different IC-values for the leaves concepts (view marked concepts). This is considered an important advantage when compared to the methods of Sanchez et al. [17] and Seco et al. [15] which provided the same IC-value for all the leaf nodes. It can, therefore, be concluded that our proposed method does not depend on and is not affected by any one parameter. A wider range of values is also provided due to the integration of two topological parameters (depth and descendants) in the specificity computation process.

5. Results and discussion

The evaluation of semantic similarity measure is expressed using the correlation coefficients (Eqs. (23) and (24)) between the human judgments assigned to a set of biomedical concepts into the datasets in Table 3 and the values calculated automatically by the proposed measure. Experiments are performed by applying some taxonomical measures already cited in the benchmarks presented in Table 3 using MeSH.

5.1. Datasets

Table 3 contains the benchmarks formed by human judgments and used to assess different measures for semantic similarity purposes.

The first dataset **MeSH1** [9] is created in collaboration with experts from the Mayo Clinic and consists of a set of word pairs referring to general medical disorders. The similarity of each concept pair was assessed by a group of 9 medical coders and 3 physicians who were aware of the notion of semantic similarity. A final

Table 3
Biomedical datasets used in evaluation of semantic similarity task.

Dataset	Year	# Pairs	Ref.
MeSH1	2007	30	[9]
MeSH2	2005	36	[35]
UMNSRS	2010	241	[43]

set of 30 word pairs with the averaged similarity measures provided by experts in a continuous scale between 1 and 4 were obtained. The second biomedical benchmark **MeSH2**, proposed by [35], consisted of a set of 36 word pairs extracted from the MeSH repository. The similarity between word pairs was also assessed by 8 medical experts from 0 (non-similar) to 1 (synonyms).

The reference standard used in our experiments was based on a set of medical pairs of terms created specifically for testing automated measures of semantic relatedness as part of a different study [43]. The pairs of terms were initially compiled by selecting all concepts from the UMLS with one of three semantic types: disorders, symptoms and drugs. Only the concepts with entry terms containing at least one single-word term were then selected for potential differences in similarity and relatedness responses. Five medical residents at the University of Minnesota Medical School were invited to participate in this study for a modest monetary compensation. They were presented with 724 medical pairs of terms on a touch-sensitive computer screen and asked to indicate the degree of relatedness between them on a continuous scale by touching a touch-sensitive bar at the bottom of the screen. In these experiments, we exploited a subset formed by 241 pairs (UMNSRS: University of Minnesota miNneapolis Semantic Relatedness/Similarity) existing in MeSH.

5.2. Evaluation metrics

The comparison between the values provided by a measure and human judgments is particularly based on correlation coefficients.

Pearson coefficient: the Pearson product-moment correlation coefficient r can be employed as an evaluation metric. It indicates how well the results of a measure resemble human judgments, where a value of 0 means no correlation and 1 means perfect correlation. The Pearson correlation coefficient r is calculated as follows:

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2)(\sum x_i)^2}\sqrt{n(\sum y_i^2)(\sum y_i)^2}} \quad (23)$$

where x_i refers to the ith element in the list of human judgments, y_i to the corresponding ith element in the list of semantic similarity computed values, and n to the number of word pairs.

Spearman coefficient: it is used to correlate word pair rankings. The ranking produced on the basis of the measure is compared to the one produced on the basis of human judgments. The quality of such ranking is quantified by the Spearman rank order correlation coefficient (ρ). For example, when a semantic similarity measure outputs a numerical value within the range [0; 1] instead of ranks, the raw values are converted into ranks. The parameter d_i is the difference between the ranks of x_i and y_i.

Table 4
Results of the proposed IC-computing method used with Lin similarity measure and applied on the MeSH1, MeSH2 and UMNSRS benchmarks. Performances are expressed using the correlation coefficients Pearson (r) and Spearman (ρ).

		Hyponyms' subgraph quantification				
		Eq. (14) Seco et al.	Eq. (15) Sanchez et al.	Eq. (16) Meng et al.	Eq. (17) Wang et al.	Eq. (18) Taieb et al.
MeSH1 (physicians)						
depthmax	r	0.766	0.758	0.765	0.765	0.756
	ρ	**0.652**	0.650	0.648	0.647	0.634
depthmin	r	**0.767**	0.760	0.760	0.766	0.758
	ρ	**0.652**	0.650	0.649	0.647	0.634
$depth_{WH}$	r	0.762	0.764	0.759	0.760	0.745
	ρ	0.642	0.643	0.642	0.644	0.643
MeSH1 (coders)						
depthmax	r	0.868	0.871	0.867	0.868	0.849
	ρ	**0.682**	0.679	0.673	0.673	0.667
depthmin	r	0.870	**0.873**	0.869	0.870	0.851
	ρ	**0.682**	**0.682**	0.673	0.673	0.667
$depth_{WH}$	r	0.804	0.813	0.799	0.801	0.785
	ρ	0.668	0.669	0.665	0.669	0.680
MeSH2						
depthmax	r	0.732	0.732	0.731	0.731	**0.746**
	ρ	0.724	0.727	0.728	0.728	0.740
depthmin	r	0.732	0.732	0.731	0.731	**0.746**
	ρ	0.731	0.735	0.727	0.725	0.742
$depth_{WH}$	r	0.731	0.729	0.724	0.725	0.741
	ρ	0.725	0.717	0.720	0.719	**0.753**
UMNSRS						
depthmax	r	**0.772**	**0.772**	0.768	0.769	0.746
	ρ	0.634	0.637	0.633	0.634	0.611
depthmin	r	0.769	0.769	0.765	0.766	0.744
	ρ	0.636	**0.638**	0.634	0.635	0.613
$depth_{WH}$	r	0.737	0.742	0.728	0.731	0.707
	ρ	0.619	0.618	0.619	0.619	0.591

The bold values represent the highest correlations values.

Table 5
Results of IC approaches combining IC-computing methods and IC-based measures. IC-based measures are Lin [31] (*Lin*), Resnik [1] (*Res*), Pirro [23] (*Pir*), Meng [33] (*Men*) and Jiang and Conrath [30] (JC). Performances are expressed using the correlation coefficients Pearson (r) and Spearman (ρ).

		Lin	Res	Pir	Men	JC
MeSH1(Physicians)						
Seco et al. [15]	r	0.678	0.669	0.251	0.687	−0.665
	ρ	0.624	0.585	0.205	0.624	−0.612
Sebti and Barfroush [14]	r	0.571	0.475	0.429	0.315	−0.389
	ρ	0.588	0.472	0.312	0.588	−0.275
Zhou et al. [16]	r	0.637	0.627	0.662	0.671	−0.645
	ρ	0.594	0.565	0.631	0.594	−0.616
Sanchez et al. [17]	r	0.664	0.656	0.670	0.683	−0.647
	ρ	0.627	0.584	0.631	0.627	−0.615
Meng et al. [13]	r	0.696	0.691	0.690	0.699	−0.654
	ρ	0.623	0.593	0.638	0.623	−0.619
AsGIC	r	**0.766**	0.625	0.571	0.740	−0.523
	ρ	**0.652**	0.520	0.617	**0.652**	−0.578
MeSH1(Coders)						
Seco et al. [15]	r	0.789	0.766	0.054	0.832	−0.660
	ρ	0.676	0.650	0.064	0.676	−0.533
Sebti and Barfroush [14]	r	0.594	0.467	0.436	0.371	−0.402
	ρ	0.576	0.562	0.310	0.576	−0.257
Zhou et al. [16]	r	0.698	0.692	0.670	0.767	−0.609
	ρ	0.629	0.613	0.586	0.629	−0.537
Sanchez et al. [17]	r	0.750	0.739	0.696	0.804	−0.648
	ρ	0.678	0.650	0.581	0.678	−0.539
Meng et al. [13]	r	0.820	0.813	0.708	0.859	−0.633
	ρ	0.666	0.655	0.574	0.666	−0.548
AsGIC	r	0.868	0.742	0.489	**0.879**	−0.424
	ρ	**0.682**	0.600	0.509	**0.682**	−0.466
MeSH2						
Seco et al. [15]	r	0.735	0.708	−0.184	0.759	−0.756
	ρ	0.753	0.689	−0.247	0.753	−0.753
Sebti and Barfroush [14]	r	0.084	0.287	0.343	0.220	−0.357
	ρ	−0.037	0.260	0.251	−0.037	−0.248
Zhou et al. [16]	r	0.707	0.686	0.721	0.738	−0.735
	ρ	**0.782**	0.683	0.746	**0.782**	−0.774
Sanchez et al. [17]	r	0.694	0.683	0.707	0.723	−0.717
	ρ	0.755	0.690	0.737	0.755	−0.751
Meng et al. [13]	r	0.753	0.710	0.776	**0.773**	−0.792
	ρ	0.780	0.682	0.754	0.780	−0.785
AsGIC	r	0.732	0.414	0.577	0.725	−0.502
	ρ	0.724	0.632	0.651	0.727	−0.670
UMNSRS						
Seco et al. [15]	r	0.697	0.695	−0.343	0.698	−0.687
	ρ	0.597	0.586	−0.179	0.597	−0.600
Sebti and Barfroush [14]	r	0.260	0.491	0.489	0.190	−0.486
	ρ	0.226	0.511	0.466	0.226	−0.455
Zhou et al. [16]	r	0.690	0.693	0.693	0.712	−0.687
	ρ	0.604	0.589	0.606	0.604	−0.606
Sanchez et al. [17]	r	0.691	0.689	0.692	0.706	−0.692
	ρ	0.600	0.586	0.603	0.600	−0.602
Meng et al. [13]	r	0.724	0.715	0.721	0.723	−0.697
	ρ	0.610	0.589	0.608	0.610	−0.601
AsGIC	r	**0.772**	0.422	0.379	0.754	−0.187
	ρ	**0.634**	0.567	0.438	0.634	−0.298

The bold values represent the highest correlations values.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (24)$$

5.3. Results

The correlations between the human judgments of the exploited biomedical benchmarks and the computed values are expressed using the Pearson (r) and Spearman (ρ) correlation coefficients. Experiments are performed in two steps:

- The first step examines the best way for quantifying the used topological parameters in the proposed IC computing method (AsGIC): the depth (*depthmax*, *depthmin* or $depth_{WH}$) and the descendants' subgraph (view Eqs. (14)–(18)). Results in Table 4 are provided using the intrinsic IC computing method with the Lin similarity measure.
- The second step is a comparison between our IC computing method (using the *depthmax* for the depth and the cardinality of descendants for the descendants' subgraph quantification) and other intrinsic methods. This comparison is performed through the different similarity measures exploiting the IC-values of two target concepts c_1 and c_2 with their lowest common subsumer. The results are presented in the Table 5.

Table 4 presents the correlation values obtained for the Pearson (r) and Spearman (ρ) coefficient values when applying our IC computing method (AsGIC) coupled with the Lin similarity measure on different biomedical benchmarks. During the experiments, we varied the ways for expressing the depth and the descendants' subgraph taxonomical parameters. We noted that the *depthmax* and *depthmin* have very close performances for all the datasets. As for the $depth_{WH}$, it shows good Spearman correlation with the dataset MeSH2 ($\rho = 0.753$) when coupled with the probabilities distribution method [36] for quantifying the descendants' subgraph.

From this first step, we conclude that our proposed computing IC method performs well using the depth parameter: *depthmax* or *depthmin*. The nearest performances in relation to *depthmax* and *depthmin* are due to the fact that 61.16% of biomedical concepts have the same depth in MeSH taxonomy. Concerning the quantification of the descendants' subgraph, the ways that can be exploited are the cardinality of the descendants (Eq. (14)), the leaves' cardinality (Eq. (15)) or the probabilities distribution (Eq. (18)). Moreover, the quantification of the descendants' subgraph using the depth of each concept (Eqs. (16) and (17)) does not show good results due to the fact that 66.75% of concepts pertaining to the MeSH taxonomy are leaves' nodes. So, they do not have developed descendants' subgraph. For this reason, using the depth of each concept for quantifying the descendants' subgraph does not display a very influential effect. We, therefore, choose to exploit the *depthmax* and the descendants' cardinality to compare the AsGIC with other intrinsic IC methods through the different IC-based similarity measures (Table 5).

Table 5 shows that our IC computing method (integrating *depthmax* and the *|descendants(c)|*) coupled with the Lin similarity measure or the Meng et al. measure provides the highest correlations. This is explained by the fact that each biomedical concept from the pair is modeled by its ancestors' subgraph. Accordingly, the common features which represent the semantic similarity are modeled by the ancestors' subgraph of the LCS ($IC(LCS(c_1, c_2))$). Therefore, the most adequate similarity measure for our proposed computing method is the Lin measure [31]. Moreover, our proposal outperforms all other intrinsic IC computing methods for all datasets except for MeSH2. For example, with the MeSH1 benchmark (coders), we obtained an excellent correlation value $r = 0.879$. The provided semantic similarity values using the proposed IC method (*depthmax* and the descendants' cardinality) in combination with the Lin measure for each dataset are presented in Appendices A–C.

In conclusion, the results from the comparative analysis show that our IC computing method reaches the highest correlations with Lin [31] or Meng et al. [33] measures. It is worth noting that the Meng et al. (Eq. (6)) measure is derived from the Lin measure (Eq. (3)). Both measures have, therefore, yielded into very close results. The Lin measure corresponds to the Dice coefficient which tries to quantify the overlapped part between the ancestors' subgraphs of the two biomedical concepts c_1 and c_2 concerned by the calculation of the semantic similarity. The overlapped part between the ancestors' subgraphs is the ancestors' subgraph of the lowest common subsumer which is represented by the $IC(LCS(c_1, c_2))$.

6. Conclusion and future work

In this paper, we propose a new intrinsic computing method for that quantifies the Information Content (IC) of a biomedical concept. The IC is modeled by the ancestors' subgraph due to the high frequency of the multiple inheritances in the biomedical knowledge resources. The IC calculation process is based on the specificity of each ancestor. This specificity is quantified using the topological parameters extracted from the "*is a*" hierarchy, namely the depth and descendants' subgraph. In this paper, we have also described in detail the different ways in which the depth and descendants' subgraph can be expressed and quantified. We focused on the IC-based semantic similarity measures because they can express, as accurately as possible, the semantics contents assigned to a concept pertaining to an "*is a*" knowledge structure. Experimental assays were performed using well-known benchmarks, MeSH1, MeSH2 and UMNSRS. The results showed that our proposal outperformed other computing IC methods with all the benchmarks used, except for MeSH2. The specificity of hypernyms is well expressed for the depth side, *depthmax* and *depthmin*, as well as for the descendants' subgraph side. It can be expressed using the cardinality (Eq. (14)), leaves (Eq. (15)), or depth distribution (Eq. (18)). The best correlations are obtained when our proposed method is used in combination with the Lin similarity measure for estimating the similarity degree through the common features represented by the ancestors' subgraph of the lowest common subsumer. Considering the promising results yielded by our proposed IC-computing method, further studies, some of which are currently underway in our laboratories, are needed to explore its feasibility and potential in the semantic grouping of clinical terms.

Conflict of interest

None declared.

Acknowledgments

The authors would like to express their gratitude to Mr. Anouar Smaoui from the English Language Unit at the Faculty of Sciences of Sfax for his valuable proofreading and language editing services.

Appendix A

Semantic similarity computed for the dataset MeSH1 with the proposed IC computing method using the *depthmax* and the cardinality of descendants, and Lin similarity measure.

	Term 1	DUI	Term 2	DUI	Value
1	Renal failure	*D051437*	Kidney failure	*D051437*	1.000
2	Abortion	*D000022*	Miscarriage	*D000022*	1.000
3	Delusion	*D003702*	Schizophrenia	*D019967*	0.108
4	Metastasis	*D009362*	Adenocarcinoma	*D000230*	0.039
5	Calcification	*D055956*	Stenosis	*D016893*	0.439
6	Mitral stenosis	*D008946*	Atrial fibrillation	*D001281*	0.212
7	Rheumatoid arthritis	*D001172*	Lupus	*D008178*	0.085
8	Carpal tunnel syndrome	*D002349*	Osteoarthritis	*D010003*	0.025
9	Diabetes mellitus	*D003920*	Hypertension	*D009798*	0.104
10	Acne	*D054506*	Syringe	*D013594*	0.000
11	Antibiotic	*D000903*	Allergy	*D057985*	0.204
12	Multiple sclerosis	*D020529*	Psychosis	*D000425*	0.000
13	Appendicitis	*D001064*	Osteoporosis	*D010024*	0.054
14	Depression	*D019052*	Cellulitis	*D002481*	0.037
15	Hyperlipidemia	*D006949*	Metastasis	*D009362*	0.051
16	Heart	*D006330*	Myocardium	*D056830*	0.661
17	Stroke	*D020521*	Infarct	*D020520*	0.526
18	Diarrhea	*D003967*	Stomach cramps	*D003085*	0.000
19	Xerostomia	*D014987*	Alcoholic cirrhosis	*D008104*	0.033
20	Lymphoid hyperplasia	*D000796*	Laryngeal cancer	*D007822*	0.023
21	Varicose vein	*D014648*	Entire knee meniscus	*D019645*	0.000
22	Chronic obstructive pulmonary disease	*D029424*	Lung infiltrates	*D055370*	0.000
23	Cortisone	*D003348*	Total knee replacement	*D019645*	0.000
24	Congestive heart failure	*D006333*	Pulmonary edema	*D011654*	0.575
25	Rectal polyp	*D051517*	Aorta	*D001011*	0.000
26	Peptic ulcer disease	*D013276*	Myopia	*D009216*	0.000

(*continued on next page*)

Appendix A (*continued*)

	Term 1	DUI	Term 2	DUI	Value
27	Pulmonary embolus	*D056824*	Myocardial infarction	*D056988*	0.000
28	Pulmonary fibrosis	*D011658*	Lung cancer	*D008175*	0.063
29	Cholangiocarcinoma	*D018281*	Colonoscopy	*D023881*	0.000
30	Brain tumor	*D001932*	Intracranial hemorrhage	*D020300*	0.318
			Physicians	r	**0.766**
				ρ	**0.652**
			Coders	r	**0.868**
				ρ	**0.682**

The bold values represent the results.

Appendix B

Semantic similarity computed for the dataset MeSH2 with the proposed IC computing method using the *depthmax* and the cardinality of descendants, and Lin similarity measure.

	Term 1	DUI	Term 2	DUI	Value
1	Anemia	*D000740*	Appendicitis	*D001064*	0.059
2	Dementia	*D003704*	Atopic Dermatitis	*D003876*	0.040
3	Osteoporosis	*D010024*	Patent Ductus Arteriosus	*D004374*	0.057
4	Sinusitis	*D012852*	Mental Retardation	*D008607*	0.033
5	Hypertension	*D006977*	Kidney Failure	*D051437*	0.846
6	Hyperlipidemia	*D006949*	Hyperkalemia	*D006947*	0.111
7	Hypothyroidism	*D007037*	Hyperthyroidism	*D006980*	0.301
8	Sarcoidosis	*D017565*	Tuberculosis	*D012830*	0.052
9	Asthma	*D001249*	Pneumonia	*D018549*	0.048
10	Lactose Intolerance	*D007787*	Irritable Bowel Syndrome	*D043183*	0.347
11	Urinary Tract Infection	*D014552*	Pyelonephritis	*D011704*	0.190
12	Psychology	*D011584*	Cognitive Science	*D019336*	0.994
13	Adenovirus	*D062705*	Rotavirus	*D022243*	0.961
14	Migraine	*D008881*	Headache	*D051270*	0.821
15	Hepatitis B	*D006510*	Hepatitis C	*D018937*	0.918
16	Carcinoma	*D002284*	Neoplasm	*D009374*	0.993
17	Pulmonary Stenosis	*D011666*	Aortic Stenosis	*D001024*	0.183
18	Breast Feeding	*D001942*	Lactation	*D007774*	0.010
19	Pain	*D010146*	Ache	*D010146*	1.000
20	Measles	*D008457*	Rubeola	*D008457*	1.000
21	Down Syndrome	*D004314*	Trisomy 21	*D004314*	1.000
22	Diabetic Nephropathy	*D003928*	Diabetes Mellitus	*D003920*	0.458
23	Neonatal Jaundice	*D007567*	Sepsis	*D018805*	0.045
24	Vaccines	*D014612*	Immunity	*D007109*	0.000
25	Amino Acid Sequence	*D000595*	Antibacterial Agents	*D000900*	0.000
26	Acq. Immunno. Syndrome	*D000163*	Congenital Heart Defects	*D006330*	0.000
27	Bacterial Pneumonia	*D018410*	Malaria	*D008288*	0.057
28	Otitis Media	*D010033*	Infantile Colic	*D003085*	0.009
29	Meningitis	*D008581*	Tricuspid Atresia	*D018785*	0.048
30	Anemia	*D018798*	Deficiency Anemia	*D018798*	1.000
31	Antibiotics	*D000900*	Antibacterial Agents	*D000900*	1.000
32	Seizures	*D012640*	Convulsions	*D012640*	1.000
33	Failure to Thrive	*D005183*	Malnutrition	*D044342*	0.121
34	Malnutrition	*D044342*	Nutritional Deficiency	*D044342*	1.000
35	Chicken Pox	*D002644*	Varicella	*D002644*	1.000
36	Myocardial Ischemia	*D017202*	Myocardial Infarction	*D009203*	0.664
				r	**0.732**
				ρ	**0.724**

The bold values represent the results.

Appendix C

Semantic similarity computed for the dataset UMNSRS with the proposed IC computing method using the *depthmax* and the cardinality of descendants, and Lin similarity measure.

	Term 1	DUI	Term 2	DUI	Value
1	Iron	*D007501*	Iron	*D007501*	1.000
2	Medrol	*D008775*	Prednisolone	*D008775*	1.000
3	Convulsion	*D012640*	Epilepsy	*D004827*	0.695
4	Emaciation	*D004614*	Cachexia	*D002100*	0.960
5	Dizziness	*D004244*	Vertigo	*D014717*	0.496
6	Mycosis	*D003047*	Histoplasmosis	*D006660*	0.772
7	Enalapril	*D004656*	Lisinopril	*D017706*	0.575
8	Xanax	*D000525*	Ativan	*D008140*	0.662
9	Ethanol	*D000432*	Alcohol	*D000438*	0.977
10	Ampicillin	*D000667*	Amoxil	*D000658*	0.974
11	Arthralgia	*D018771*	Pain	*D010146*	0.993
12	Amantadine	*D000547*	Tamiflu	*D053139*	0.478
13	Cefaclor	*D002433*	Cefoxitin	*D002440*	0.919
14	Hepatitis	*D006519*	Cirrhosis	*D008104*	0.717
15	Carsickness	*D009041*	Nausea	*D009325*	0.593
16	Weakness	*D018908*	Paresis	*D010291*	0.688
17	Candidiasis	*D002178*	Mycosis	*D002862*	0.864
18	Allopurinol	*D000493*	Colchicine	*D003078*	0.455
19	Mycosis	*D003047*	Coccidioses	*D003048*	0.370
20	Syphilis	*D013587*	Gonorrhea	*D006069*	0.502
21	Mycosis	*D001759*	Blastomycoses	*D001759*	1.000
22	Torticollis	*D014103*	Spasm	*D013035*	0.340
23	Myositis	*D009220*	Myopathy	*D009135*	0.875
24	Penicillin	*D010406*	Cefazolin	*D002437*	0.845
25	Rhonchi	*D012135*	Rales	*D012135*	1.000
26	Encephalitis	*D004660*	Headache	*D020773*	0.409
27	Dyspnea	*D004417*	Cyanosis	*D003490*	0.519
28	Nausea	*D020250*	Vomiting	*D020250*	1.000
29	Encephalitis	*D004660*	Meningitis	*D008587*	0.370
30	Bronchitis	*D001991*	Pneumonia	*D018549*	0.404
31	Dermatomyositis	*D003882*	Myopathy	*D009135*	0.342
32	Hematemesis	*D006396*	Vomiting	*D014839*	0.477
33	Avitaminoses	*D001361*	Starvation	*D013217*	0.393
34	Coccidioidomycosis	*D003047*	Histoplasmosis	*D006660*	0.772
35	Clonus	*D009207*	Spasm	*D013035*	0.508
36	Photosensitization	*D010787*	Dermatitis	*D003872*	0.722
37	Hepatitis	*D006505*	Hepatomegaly	*D006529*	0.201
38	Hemoptysis	*D006469*	Hematemesis	*D006396*	0.272
39	Pallor	*D010167*	Iron	*D019189*	0.325
40	Influenza	*D044135*	Pneumoniae	*D016977*	0.590
41	Anemia	*D000740*	Reticulocytosis	*D045262*	0.230
42	Catatonia	*D002389*	Lethargy	*D053609*	0.871
43	Sleeplessness	*D007319*	Agitation	*D011595*	0.230
44	Meningitis	*D016920*	Tuberculosis	*D014390*	0.470
45	Seasickness	*D009041*	Nausea	*D009325*	0.593
46	Cardiomyopathy	*D009202*	Angina	*D008158*	0.370
47	Hemochromatosis	*D006432*	Arthritis	*D015210*	0.359
48	Glomerulosclerosis	*D005923*	Proteinuria	*D011507*	0.252
49	Anosmia	*D000857*	Aphonia	*D001044*	0.658
50	Vancocin	*D014640*	Glucotrol	*D005913*	0.187
51	Dyslipidemia	*D050171*	Hyperlipidemia	*D006949*	0.626
52	Dyslipidemia	*D050171*	Angina	*D008158*	0.133
53	Flatulence	*D005414*	Halitosis	*D006209*	0.788
54	Diarrhea	*D000930*	Colchicine	*D003078*	0.272
55	Coccidioses	*D003048*	Meningitis	*D008581*	0.131
56	Fibrillation	*D001281*	Angina	*D008158*	0.179

(*continued on next page*)

Appendix C (*continued*)

	Term 1	DUI	Term 2	DUI	Value
57	Fibrillation	*D001281*	Cardiomyopathies	*D009202*	0.240
58	Activase	*D010959*	Streptase	*D013300*	0.871
59	Earache	*D004433*	Hyperacusis	*D012001*	0.134
60	Neuropathy	*D018917*	Insulin	*D006946*	0.140
61	Arteriosclerosis	*D001161*	Ischemias	*D007511*	0.118
62	Pancreatitis	*D010195*	Insulin	*D006946*	0.163
63	Anoxemia	*D000860*	Seizures	*D012640*	0.114
64	Syncope	*D013575*	Weakness	*D018908*	0.288
65	Cirrhosis	*D008103*	Anemia	*D000740*	0.261
66	Fibrillation	*D001281*	Angina	*D008158*	0.179
67	Arthritis	*D001168*	Amyloidoses	*D000686*	0.085
68	Smallpox	*D012899*	Vaccinia	*D014615*	0.666
69	Cardiomyopathy	*D009202*	Rales	*D030341*	0.347
70	Heparin	*D006493*	Protamine Sulfate	*D011479*	0.193
71	Rales	*D030341*	Cyanosis	*D003490*	0.417
72	comatose	*D003128*	Anoxemia	*D000860*	0.051
73	Adenitis	*D008199*	Stridor	*D012135*	0.135
74	Lasix	*D005665*	Mannitol	*D008353*	0.062
75	Cataracts	*D002386*	Diabetes	*D003920*	0.085
76	Stridor	*D012135*	Snoring	*D012913*	0.634
77	Cisplatin	*D002945*	Zofran	*D017294*	0.102
78	Angina	*D008158*	Dyspnea	*D004417*	0.244
79	Anemia	*D051856*	Coumadin	*D014859*	0.068
80	Pneumonia	*D054988*	Cyanosis	*D003490*	0.190
81	Uremias	*D014511*	Nausea	*D009325*	0.066
82	Dysuria	*D053159*	Cipro	*D054139*	0.049
83	Hyperacusis	*D12001*	Bleomycin	*D001761*	0.000
84	Hypothyroidism	*D007037*	Infertility	*D007246*	0.128
85	Dyspnea	*D004417*	Agitation	*D011595*	0.072
86	Deafness	*D003638*	Ataxia	*D001259*	0.241
87	Blastomycoses	*D001759*	Seizures	*D003294*	0.044
88	Codeine	*D003061*	Narcan	*D009270*	0.732
89	Dyspnea	*D004417*	Narcan	*D009270*	0.000
90	Anovulation	*D000858*	Emaciation	*D004614*	0.003
91	Otitis	*D010031*	Meningism	*D008580*	0.114
92	Epilepsy	*D004827*	Alcohol	*D020270*	0.255
93	Comatose	*D003128*	Hepatitis	*D006505*	0.024
94	Hemiplegia	*D006429*	Headache	*D006261*	0.150
95	Uremias	*D014511*	Pyorrhea	*D010510*	0.062
96	Dehydration	*D003681*	Candidiases	*D002177*	0.222
97	Hemoptysis	*D006469*	Protamine Sulfate	*D011479*	0.000
98	Virilism	*D014770*	Drooling	*D012798*	0.260
99	Meningism	*D008580*	Hyperesthesia	*D006941*	0.664
100	Dyslipidemia	*D050171*	Activase	*D006949*	0.000
101	Pallor	*D010167*	Albumin	*D050010*	0.126
102	Hypoproteinemia	*D007019*	Hunger	*D006815*	0.000
103	Flushing	*D005483*	Iron	*D019189*	0.472
104	Arteriosclerosis	*D001161*	Cholestyramine	*D002792*	0.000
105	Diarrhea	*D000930*	Cefaclor	*D002433*	0.013
106	Epilepsy	*D004827*	Ischemia	*D002545*	0.627
107	Hemoglobinopathy	*D006453*	Pain	*D046788*	0.360
108	Uremias	*D014511*	Emaciation	*D004614*	0.004
109	Cirrhosis	*D005355*	Zocor	*D019821*	0.000
110	Anoxemia	*D000860*	Ataxia	*D001259*	0.098
111	Hyperthyroidism	*D006980*	Osteoporosis	*D010024*	0.162
112	Snoring	*D012913*	Stridor	*D012135*	0.634
113	Narcan	*D009270*	Duragesic	*D005283*	0.031
114	Anoxemia	*D000860*	Propofol	*D015742*	0.000
115	Neuropathy	*D018917*	Constipation	*D003248*	0.179
116	Herpes	*D006566*	Cholestasis	*D002779*	0.071

Appendix C (*continued*)

	Term 1	DUI	Term 2	DUI	Value
117	Thrombophilias	*D019851*	Hemoptysis	*D006469*	0.184
118	Ischemias	*D007511*	Drooling	*D012798*	0.264
119	Erythema	*D004890*	Dilantin	*D010672*	0.000
120	Seasickness	*D009041*	Ethanol	*D000431*	0.000
121	Catatonia	*D002389*	Narcan	*D009270*	0.000
122	Hemophilia	*D006467*	Thromboembolism	*D013923*	0.024
123	Camelpox	*D018155*	Echinacea	*D020900*	0.010
124	Bacteremia	*D016470*	Amauroses	*D001766*	0.019
125	Septicemia	*D018805*	Dyspnea	*D004417*	0.034
126	Hyperacusis	*D012001*	Aphonia	*D001044*	0.444
127	Erythromycin	*D004917*	Lidocaine	*D008012*	0.053
128	Reticulocytosis	*D045262*	Hematemesis	*D006396*	0.141
129	Giardiasis	*D005873*	Avitaminoses	*D001361*	0.026
130	Leukopenia	*D007970*	Penicillin	*D010406*	0.000
131	Spasm	*D013035*	Lasix	*D005665*	0.000
132	Headache	*D020773*	Flushing	*D005483*	0.108
133	Diabetes	*D003920*	Psoriasis	*D011565*	0.083
134	Pallor	*D010167*	Aspirin	*D055963*	0.031
135	Epilepsy	*D004827*	Anosmia	*D000857*	0.057
136	Overnutrition	*D044343*	Seizures	*D003294*	0.093
137	Drooling	*D012798*	Chills	*D023341*	0.262
138	Goiter	*D006042*	Scleroderma	*D012594*	0.253
139	Uremias	*D014511*	Motrin	*D007052*	0.000
140	Cataracts	*D002386*	Insulin	*D006946*	0.159
141	Gastroenteritis	*D005759*	Cromolyn	*D004205*	0.000
142	Candidiasis	*D002177*	Medrol	*D008775*	0.000
143	Coccidioses	*D003048*	Medrol	*D008775*	0.000
144	Rheumatism	*D012216*	Sleeplessness	*D007319*	0.028
145	Dyspnea	*D004417*	Penicillin	*D010406*	0.000
146	Ketonuria	*D007662*	Folic Acid	*D005494*	0.019
147	Hemochromatosis	*D006432*	Anosmia	*D000857*	0.030
148	Albumin	*D010047*	Heparin	*D015844*	0.281
149	Vasculitis	*D020293*	Convulsion	*D012640*	0.399
150	Nystagmus	*D009759*	Incontinence	*D005242*	0.068
151	Cyanosis	*D003490*	Folic Acid	*D005494*	0.025
152	Vasculitis	*D014657*	Hemiplegia	*D006429*	0.097
153	Hernia	*D006547*	Goiter	*D006042*	0.236
154	Nocturia	*D053158*	Bacitracin	*D001414*	0.000
155	Gastroenteritis	*D005759*	Insulin	*D006946*	0.165
156	Fibrillation	*D001281*	Halitosis	*D006209*	0.154
157	Infertility	*D007246*	Fibrillation	*D001281*	0.092
158	Erythromycin	*D004917*	Allopurinol	*D000493*	0.061
159	Torticollis	*D014103*	Erythromycin	*D004917*	0.000
160	Leukopenia	*D007970*	Cefaclor	*D002433*	0.000
161	Pancreatitis	*D010195*	Cataract	*D002386*	0.192
162	Aneurysm	*D000783*	Silvadene	*D012837*	0.000
163	Bleomycin	*D001761*	Antabuse	*D004221*	0.022
164	Seizures	*D012640*	Insulin	*D007331*	0.204
165	Glucotrol	*D005913*	Motrin	*D007052*	0.134
166	Thrombocytopenia	*D013921*	Heartburn	*D006356*	0.113
167	Goiter	*D006042*	Vasculitis	*D014657*	0.220
168	Chills	*D023341*	Snoring	*D012913*	0.144
169	Blepharospasm	*D001764*	Bedwetting	*D053206*	0.037
170	Starvation	*D013217*	Cisplatin	*D002945*	0.000
171	Propranolol	*D011433*	Bactroban	*D016712*	0.027
172	Dyslipidemia	*D050171*	Constipation	*D003248*	0.119
173	Erythema	*D004890*	Tremor	*D020329*	0.292
174	Hemochromatosis	*D006432*	Polyuria	*D011141*	0.030
175	Ataxia	*D001259*	Constipation	*D003248*	0.100
176	Wellbutrin	*D016642*	Endep	*D000639*	0.042
177	Cataracts	*D002386*	Pancreatitis	*D010195*	0.192

(*continued on next page*)

Appendix C (*continued*)

	Term 1	DUI	Term 2	DUI	Value
178	Thrombocytopenia	*D013921*	Arthritis	*D001168*	0.087
179	Prilosec	*D009853*	Glucophage	*D008687*	0.064
180	Osteoporosis	*D010024*	Cardiomyopathy	*D009202*	0.177
181	Zofran	*D017294*	Ipecac Syrup	*D007486*	0.087
182	Silvadene	*D012837*	Ketamine	*D007649*	0.026
183	Anosmia	*D000857*	Digoxin	*D004077*	0.000
184	Histoplasmosis	*D006660*	Prednisolone	*D011239*	0.000
185	Asthma	*D001249*	Urolithiasis	*D052878*	0.034
186	Proteinuria	*D011507*	Rogaine	*D008914*	0.000
187	Photosensitization	*D010787*	Aphonia	*D001044*	0.077
188	Peritonitis	*D010538*	Ataxia	*D001259*	0.037
189	Glaucoma	*D057066*	Fibrillation	*D001281*	0.131
190	Plague	*D010931*	Vivarin	*D002110*	0.075
191	Glaucoma	*D057066*	Fibrillation	*D001281*	0.131
192	Cataracts	*D002386*	Glucophage	*D008687*	0.000
193	Mycosis	*D003047*	Prostatism	*D053448*	0.103
194	Angina	*D008158*	Diarrhea	*D003967*	0.279
195	Synthroid	*D013974*	Lidocaine	*D008012*	0.034
196	comatose	*D003128*	Cataracts	*D002386*	0.024
197	Cataracts	*D002386*	Narcan	*D009270*	0.000
198	Acetylcysteine	*D000111*	Adenosine	*D058892*	0.089
199	Rabies	*D011818*	Acne	*D017486*	0.061
200	Uremias	*D014511*	Aspirin	*D055963*	0.023
201	Toothache	*D014098*	Ataxia	*D001259*	0.120
202	Plague	*D010931*	Cholestyramine	*D002792*	0.017
203	Ischemias	*D007511*	Aloe Vera	*D000504*	0.000
204	Rales	*D012135*	Hyperacusis	*D012001*	0.083
205	Anosmia	*D000857*	Constipation	*D003248*	0.192
206	Exophthalmos	*D005094*	Prostatism	*D053448*	0.084
207	Bronchitis	*D001991*	Loperamide	*D008139*	0.000
208	Snoring	*D012913*	Loperamide	*D008139*	0.000
209	Cardiomyopathy	*D009202*	Tylenol	*D000082*	0.000
210	Osteoporosis	*D010024*	Thrombus	*D013927*	0.072
211	Dyslipidemia	*D050171*	Iron	*D019189*	0.242
212	Rabies	*D011819*	Silvadene	*D012837*	0.040
213	Hernias	*D006547*	Dementia	*D003704*	0.073
214	Overnutrition	*D044343*	Sandimmune	*D016572*	0.000
215	Thalassemia	*D013789*	Tremor	*D020329*	0.028
216	Aneurysm	*D000783*	Osteoporosis	*D010024*	0.182
217	Anovulation	*D000858*	Torticollis	*D014103*	0.014
218	Infertility	*D007246*	Blepharospasm	*D001764*	0.130
219	Atherosclerosis	*D050197*	Toothache	*D014098*	0.009
220	Dementia	*D003704*	Aloe Vera	*D000504*	0.000
221	Mannitol	*D008353*	Tylenol	*D000082*	0.170
222	Psoriasis	*D011565*	Meningism	*D008580*	0.101
223	Psoriasis	*D011565*	Spasm	*D001986*	0.262
224	Cataracts	*D002386*	Wheezing	*D012135*	0.111
225	Cortisone	*D013761*	Adenosine	*D058907*	0.144
226	Hemochromatosis	*D013761*	Vermox	*D058907*	0.000
227	Epilepsy	*D004827*	Hepatomegaly	*D006529*	0.086
228	Blepharospasm	*D001764*	Coumadin	*D014859*	0.000
229	Airsickness	*D009041*	Osteoporosis	*D010024*	0.193
230	Schistosomiasis	*D012552*	Cardura	*D017292*	0.000
231	Brucellosis	*D002006*	Narcan	*D009270*	0.000
232	Psoriasis	*D011565*	Coreg	*D009364*	0.199
233	Constipation	*D003248*	Calan	*D014700*	0.000
234	Hernias	*D006547*	Earache	*D004433*	0.012
235	Constipation	*D003248*	Cardizem	*D004110*	0.000
236	Psoriasis	*D011565*	Atherosclerosis	*D050197*	0.077

Appendix C (*continued*)

	Term 1	DUI	Term 2	DUI	Value
237	Hemophilia	*D006467*	Glucotrol	*D005913*	0.000
238	Seizures	*D012640*	Aloe Vera	*D000504*	0.000
239	Colitis	*D004760*	Epilepsy	*D004827*	0.053
240	Overnutrition	*D044343*	Malnutrition	*D044342*	0.128
241	Herpes	*D006566*	Hyperthyroidism	*D006980*	0.149
				r	**0.772**
				ρ	**0.634**

The bold values represent the results.

References

[1] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1998) 95–130.
[2] S. Patwardhan, S. Banerjee, T. Pedersen, Using measures of semantic relatedness for word sense disambiguation, in Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Text Processing, 2003, pp. 241–257.
[3] S. Aseervatham, Y. Bennani, Semi-structured document categorization with a semantic kernel, Pattern Recogn. 42 (9) (2009) 2067–2076.
[4] R.L. Cilibrasi, P.M.B. Vitanyi, The Google similarity distance, IEEE Trans. Knowl. Data Eng. 19 (3) (2007) 370–383.
[5] A. Budanitsky, G. Hirst, Evaluating WordNet-based measures of lexical semantic relatedness, Comput. Linguist. 32 (1) (Mar. 2006) 13–47.
[6] D. Sanchez, Domain Ontology Learning from the Web: An Unsupervised, Automatic and Domain Independent Approach, Akademikerverlag, AV, 2012.
[7] N. Ratprasartporn, J. Po, A. Cakmak, S. Bani-Ahmad, G. Ozsoyoglu, Context-based literature digital collection search, VLDB J. 18 (1) (Jan. 2009) 277–301.
[8] A. Hliaoutakis, G. Varelas, E. Voutsakis, E.G.M. Petrakis, E. Milios, Information retrieval by semantic similarity, Int. J. Seman. Web Inf. Syst. (IJSWIS), (July/September) (2006) (Special Issue of Multimedia Semantics, 2006).
[9] T. Pedersen, S.V.S. Pakhomov, S. Patwardhan, C.G. Chute, Measures of semantic similarity and relatedness in the biomedical domain, J. Biomed. Inform. 40 (3) (2007) 288–299.
[10] V. Sugumaran, V.C. Storey, Ontologies for conceptual modeling: their creation, use, and management, Data Knowl. Eng. 42 (3) (2002) 251–271.
[11] B.T. McInnes, T. Pedersen, Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs, J. Biomed. Inform. 54 (2015) 329–336.
[12] K.R. Gøeg, R. Cornet, S.K. Andersen, Clustering clinical models from local electronic health records based on semantic similarity, J. Biomed. Inform. 54 (2015) 294–304.
[13] L. Meng, J. Gu, Z. Zhou, A new model of information content based on concept's topology for measuring semantic similarity in WordNet, Int. J. Grid Distrib. Comput. 5 (3) (Sep. 2012).
[14] A. Sebti, A. A. Barfroush, A new word sense similarity measure in WordNet, in: IMCSIT, 2008, pp. 369–373.
[15] N. Seco, T. Veale, J. Hayes, An intrinsic information content metric for semantic similarity in WordNet, in: Proceedings of ECAI, vol. 4, 2004, pp. 1089–1090.
[16] Z. Zhou, Y. Wang, J. Gu, A new model of information content for semantic similarity in WordNet, in: International Conference on Future Generation Communication and Networking Symposia, vol. 3, 2008, pp. 85–89.
[17] D. Sánchez, M. Batet, D. Isern, Ontology-based information content computation, Knowl.-Based Syst. 24 (2) (2011) 297–303.
[18] D. SáNchez, M. Batet, Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective, J. Biomed. Inform. 44 (5) (2011) 749–759.
[19] M. Batet, D. Sánchez, A. Valls, K. Gibert, Semantic similarity estimation from multiple ontologies, Appl. Intell. 38 (1) (2013) 29–44.
[20] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database (Language, Speech, and Communication), illustrated ed., The MIT Press, 1998.
[21] G.B. Melton, S. Parsons, F.P. Morrison, A.S. Rothschild, M. Markatou, G. Hripcsak, Inter-patient distance metrics using SNOMED CT defining relationships, J. Biomed. Inform. 39 (6) (2006) 697–705.
[22] E.G.M. Petrakis, G. Varelas, A. Hliaoutakis, P. Raftopoulou, X-similarity: computing semantic similarity between concepts from different ontologies, J. Digital Inform. Manage., JDIM 4 (2006).
[23] G. Pirró, A semantic similarity metric combining features and intrinsic information content, Data Knowl. Eng. 68 (11) (Nov. 2009) 1289–1308.
[24] J.E. Caviedes, J.J. Cimino, Towards the development of a conceptual distance metric for the UMLS, J. Biomed. Inform. 37 (2) (Apr. 2004) 77–85.
[25] C. Ronald, Information-content-based measures for the structure of terminological systems and for data recorded using these systems, in: Proceedings of the 13th World Congress on Medical Informatics, 2010, pp. 1075–1079.
[26] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948).
[27] R. Jiang, M. Gan, X. Dou, From ontology to semantic similarity: calculation of ontology-based semantic similarity, Scient. World J. 2013 (2013).
[28] Z.Z.J. Gu, L. Meng, A review of information content metric for semantic similarity, Advan. Dig. Telev. Wirel. Multim., Commun. 331 (2012) 299–306.
[29] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, F.M. Couto, Semantic similarity in biomedical ontologies, PLoS Comput. Biol. 5 (7) (2009).
[30] J.J. Jiang, D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, CoRR (1997). cmp-lg/9709008.
[31] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, 1998, pp. 296–304.
[32] A. Tversky, Features of similarity, Psychol. Rev. 84 (4) (1977) 327–352.
[33] L. Meng, J. Gu, A new model for measuring word sense similarity in WordNet, in: Proceedings of the 4th International Conference on Advanced Communication and Networking, 2012, pp. 18–23.
[34] W.N. Francis, H. Kucera, Frequency Analysis of English Usage: Lexicon and Grammar, Houghton Mifflin, 1983.
[35] A. Hliaoutakis, Semantic Similarity Measures in the MESH Ontology and their Application to Information Retrieval on Medline, in Technical Report, Technical Univ. of Crete (TUC), Dept. of Electronic and Computer Engineering, 2005.
[36] M.A. Hadj Taieb, M. Ben Aouicha, A. Ben Hamadou, Ontology-based approach for measuring semantic similarity, Eng. Appl. Artif. Intell. 36 (November) (2014) 238–261.
[37] H. Rubenstein, J.B. Goodenough, Contextual correlates of synonymy, Commun. ACM 8 (10) (1965) 627–633.
[38] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Lang. Cognit. Process. 6 (1) (1991) 1–28.
[39] W. Kim, A.R. Aronson, W.J. Wilbur, Automatic MeSH term assignment and quality assessment, in: AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3–7, 2001, 2001.
[40] T. Wang, G. Hirst, Refining the notions of depth and density in WordNet-based semantic similarity measures, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1003–1011.
[41] M. Batet, D. Sánchez, A. Valls, An ontology-based measure to compute semantic similarity in biomedicine, J. Biomed. Inform. 44 (1) (2011) 118–125.
[42] A. Maedche, S. Staab, Comparing Ontologies – Similarity Measures and a Comparison Study, Institute AIFB, University of Karlsruhe, 2001.
[43] S. Pakhomov, B. McInnes, T. Adam, Y. Liu, T. Pedersen, G.B. Melton, Semantic similarity and relatedness between clinical terms: an experimental study, in: Annual Symposium Proceedings/AMIA Symposium, AMIA Symposium, vol. 2010, January 2010, pp. 572–576.